International Journal of Scientific Engineering and Technology Research

ISSN 2319-8885
Vol.06,Issue.14
April-2017,
Pages:2683-2689

www.ijsetr.com

Classification of Pathological Voice Types using Artificial Neural Networks Based on MFCC Features

J. SUJANAA[1], V. SRINIVASAN[2]

[1]PG Scholar, Dept of CSE, Annamalai University, Chidambaram, Tamilnadu, India, E-mail: sujanaajohn@gmail.com.
[2]Professor, Dept of CSE, Annamalai University, Chidambaram, Tamilnadu, India, E-mail: vscseau@gmail.com.

Abstract: In this paper, the detection and classification of pathological voices is performed since it is receiving an abundant attention in the recent years. The voice recordings are collected from the Saarbruecken Voice Database at 50Khz frequency rate. The various steps involved in the development of the proposed system are, i) Collection of data, ii) Feature extraction, iii) Classification of pathological voices, iv) Detection of pathological voices v) Performance analysis. The Mel Frequency Cepstral Coefficients (MFCCs) features have been extracted from the speech signals and classified into five types of categories. The goal is to classify the different types of pathological voices from the gathered voice samples. The Artificial Neural Network (ANN) classifiers like General Regression Neural Network (GRNN), Multi Layer Perceptron Neural Network (MLPNN) and Recurrent Neural Network (RNN) are used to classify the speech signals. Thus based on the performance of the classifiers, the type of pathological voice is detected.

Keywords: ANN, GRNN, MLPNN, RNN.

I. INTRODUCTION

Speech, language refer to the means of communication used by people. It is the expression of ideas and thoughts by means of articulate vocal sounds, or the faculty of thus expressing ideas and thoughts. Speech is a natural mode of communication for people. It is an important phenomenon in our day to day life. People are so comfortable with speech that we would also like to interact with our computers via speech, rather than having to resort to primitive interfaces such as keyboards and pointing devices. A speech interface would support many valuable applications, for example, telephone directory assistance, spoken database querying for novice users, "hands-busy" applications in medicine or fieldwork, office dictation devices, or even automatic voice translation into foreign languages. Such tantalizing applications have motivated research in automatic speech recognition since the 1950s. Great progress has been made so far, especially since the 1970s, using a series of engineered approaches that include template matching, knowledge engineering, and statistical modeling. Yet computers are still nowhere near the level of human performance at speech recognition, and it appears that further significant advances will require some new insights. Voice means the sound produced in a person's larynx and uttered through the mouth, as speech or song. It can be normal or abnormal type.

A. Pathological Voice

Pathology derives from the Greek pathos which means "suffering" and ology means "study of" to give us "the study of disease,"but often pathology means the disease's behavior. We also use pathology to describe abnormal conditions that aren't really diseases. Voice pathology refers to study of abnormality in human voices. We can identify the abnormal voice using the following factors such as voice that calls the attention by itself, does not meet the social or occupational needs of the people, voice that have the aberrant quality, hoarseness, breath and harshness [1].

II. PROPOSED METHODOLOGY

Fig1. shows the block diagram of the proposed work for pathological voice classification system. The first step is to collect the voice samples from the Saarbruecken Voice Database at 50khz frequency. Then extracting the MFCCs for various types of pathological voices is done. The MFCC features are given for training and testing in various neural networks. The different types of Neural Networks used for the classification are General Regression Neural Network, Multi Layer Perceptron Neural Network and Recurrent Neural Network. These neural networks help in the classification of different types of pathological voices. Also, the pathological voice classification system is developed which detects the type of voice pathology.

A. Dataset Collection

The voice samples are collected from the Saarbruecken voice database. This database has been made available freely online. It is a collection of voice recordings from more than 2000 persons, where a session is defined as a collection of, recordings of vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch. Recording of sentence like"Good morning, how are you?" are also available. In

addition, the Electro Glotto Gram (EGG) signal is also stored for each case in a separate file. The length of the audio clips with sustained vowels is 2 seconds. All recordings are sampled at 50 kHz and their resolution is 16-bit. For our experiments only files with sustained vowels and people older than 18 are used. A total of 75 voice samples such as 15 voice samples from each type of voice pathologies.

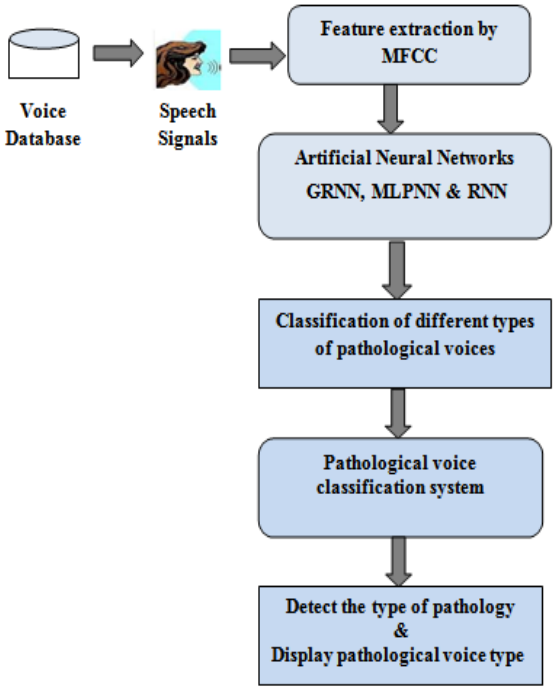


Fig.1. Block diagram of the proposed work.

B. Feature Extraction Method

Acoustic Features: Acoustic features denote the features extracted from the speaker signal. The primary and important task in voice classification is to extract the features representing the speaker information of the signal [3]. The feature extraction is the process of converting speech signals into a sequence of feature vectors carrying characteristics information about the signal. These vectors are used as basis for various types of voice analysis algorithms. It is a typical task to compute on window basis. These window based features can be considered as short time description of the signal for that particular moment in time. Difficulties arise due to limitations of the existing feature extraction techniques.

Acoustic Feature Extraction: The acoustic feature, MFCC (Mel Frequency Cepstral Coefficients) is extracted. A frame size of 20 ms and a frame shift of 10ms are used, Hence, audio signals of 0 to 2 seconds have been used to generate the feature vectors. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum") [4]. The difference between the cepstrum and the mel-frequency cepstrum is that in the MF. The frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal cepstrum. This frequency wrapping can allow for better representation of sounds [2].

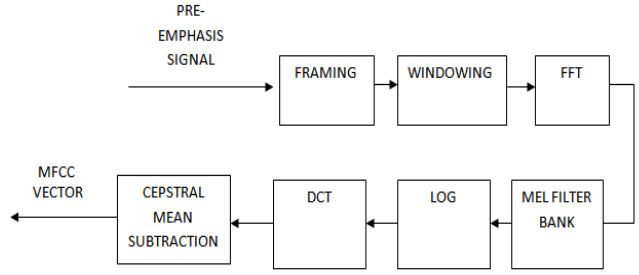


Fig.2. MFCC block diagram.

Fig.2 shows the block diagram of MFCC where the feature vector has been extracted from the pathological voice signals. MFCC employs the mel scale which is a scale of pitches which are equal in distance between one another. Mel Frequency Cepstral Coefficients (MFCC)s are features widely used in automatic speech recognition. MFCC has no nonlinear perceptual characteristics in frequency domain, which is widely used in audio analysis and recognition.

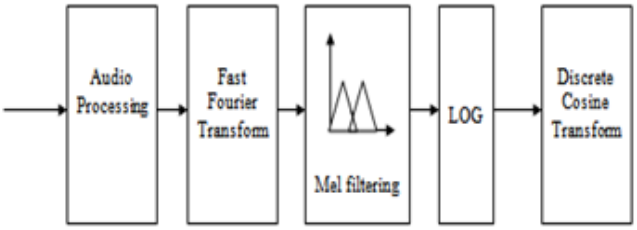


Fig.3. Mel filtering process.

Mel Frequency Cepstrum Coefficients (MFCCs): MFCCs employ the Mel scale which is a scale of pitches equal in distance from one another [4]. The normal frequency f hertz can be converted to the mel range by the following equation,

$$m = 1127.01048 \log(1 + f/700) \quad \textbf{(1)}$$

where m is the mel range. A cepstrum is the result of taking the Fourier transform of the decibel spectrum (power spectrum) as if it was a signal [2]. There is a complex cepstrum and a real cepstrum. The cepstrum can be defined mathematically as cepstrum of a signal $= FT(\log(FT(\log(signal))))$. Here the FT indicates the Fourier Transform. The real cepstrum uses the logarithm function defined real values, while the complex cepstrum uses the complex logarithm function defined for complex values. The complex cepstrum holds information about magnitude and phase of the initial spectrum, allowing the reconstruction of the signal. The real cepstrum only uses the information of the magnitude of the spectrum [5].

The cepstrum can be seen as information about rate of change in the different cepstrum bands. Usually the spectrum is first transformed using the Mel frequency bands. The result is called the MFCC's, which are used for voice identification, pitch detection etc… This is a result of the cepstrum separating the energy resulting from vocal cord phonation from the "distorted" signal formed by the rest of the vocal tract as shown in Fig.3. The human ear exhibits a

nonlinear characteristics when it comes to the perception. Hence the Mel scale takes into the account of this property. The frequency and Mel Scale coincide below 500Hz and it produces larger equal increments above it. As a result, four octaves on the hertz scale above 500 Hz are judged to rise about two octaves on the Mel scale. After the translation to the Mel frequency scale the coefficients can be evaluated. Finally the computation of MFCCs involves the windowing of the incoming audio signal. Log of the spectrum is compared and another transform is applied in order to obtain the cepstrum coefficients.

Preprocessing: To extract the features from the speaker signal, the signal must be preprocessed. Pre-emphasis, frame blocking and windowing tasks are the initial pre-processing stage before voice signal is used for the next process.

Pre-Emphasis: The higher frequencies of the audio signal are generally weak. The pre-emphasis is used to boost the energy of the high frequency signals. The digitized audio signal is put through an order digital system to spectrally flatten. To make it less susceptible to finite precision effects later in the signal processing.

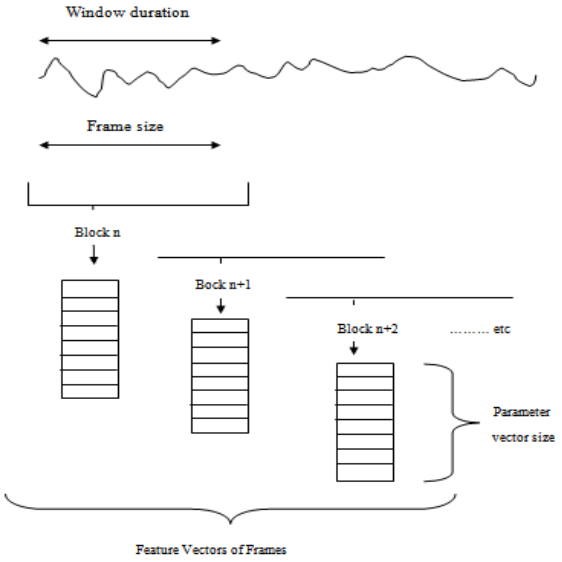


Fig.4. Frame blocking for pre-processing.

Frame Blocking: The speaker signal analysis usually assumes that the signal properties change relatively with time. This allows examination of a short time window of audio to extract parameters presumed to remain fixed for the duration of the window. Therefore the continuous audio signal is blocked into frames of N samples(frame size), adjacent frames being seperated by M samples (frame shift). Throughout this work, A frame size of 150 frames/sec, where each frame is 20 ms in duration with an overlap to 70% between adjacent frames is used throughout this work.

Windowing: Windowing is done for each individual frame so as to minimize the signal continuities at the beginning and end of each frame. The hamming window is commonly used. It tapers the samples in each window so that discriminates at the window edges are created. Fig4 shows how the continuous audio signal is divided into small frames.

C. Artificial Neural Networks

One type of network where the nodes are considered as 'artificial neurons', These are called Artificial Neural Networks (ANNs). An artificial neuron is a computational model inspired in the natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron [3]. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal through the axon. This signal might be sent to another synapse and activate other neurons.

1. General Regression Neural Network (GRNN): GRNN, as proposed by Donald F. Specht falls into the category of Probabilistic Neural Network (PNN) is discussed. This neural network like other Probabilistic Neural Networks (PNN) needs only a fraction of the training samples a Back Propagation Neural Network (BPNN) would need. The data available from measurements of an operating system will not be enough for a Back Propagation Neural Network. Therefore the use of a Probabilistic Neural Network is especially advantageous due to its ability to converge to the underlying function of the data with only few training samples available. The additional knowledge needed to get the fit in a satisfying way is relatively small and can be done without additional input by the user. This makes GRNN a very useful tool to perform predictions and comparisons of system performance in practice as shown in Fig.5.

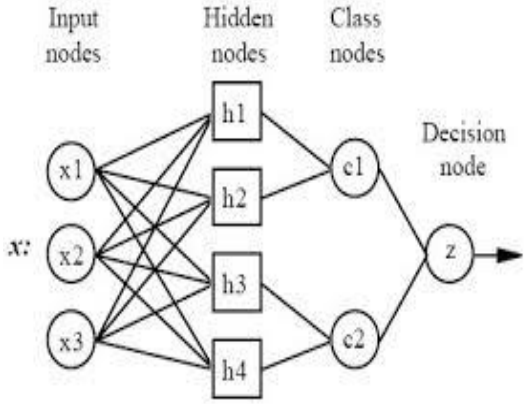


Fig.5. General structure of GRNN.

Input Layer: There is one neuron in the input layer for each predictor variable. In the case of categorical variables, N-1 neurons are used where N is the number of categories. The input neurons then feed the values to each of the neurons in the hidden layer [5].

Hidden Layer: This layer has one neuron for each case in the training data set. The neuron stores the values of the predictor variables for the case along with the target value. When presented with the x vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma value(s).

Decision Layer: The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent [15].

2. Multilayer Perceptron Neural Network (MLPNN): The most popular form of neural network architecture is the Multilayer Perceptron Neural Network (MLPNN). A multilayer perceptron:

- has any number of inputs.
- has one or more hidden layers with any number of units.
- uses linear combination functions in the input layers.
- uses generally sigmoid activation functions in the hidden layers.
- has any number of outputs with any activation functions as shown in Fig.6.
- has connections between the input layer and the first hidden layer, between the hidden layer and between the last hidden layer and the output layer.
- Given enough data, enough hidden units , enough training time, and MLP with just one hidden layer can learn to approximate virtually any function to any degree of accuracy(A statistical analogy is approximating a function with n^{th} order polynomials). For this reason, MPLs are known as universal approximators and can be used when we have little prior knowledge of the relationship between inputs and targets. Although one hidden layer is always sufficient provided we have enough data, there are situations where a network with two or more hidden layers may require fewer hidden units and weights, than a network with one hidden layer. So using extra hidden layers sometimes can improve generalization.

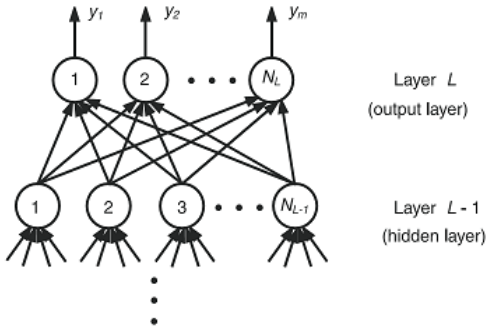


Fig.6. General structure of MLPNN.

Recurrent Neural Network (RNN): The fundamental feature of a Recurrent Neural Network (RNN) is that the network contains at least one feed-back connection, so the activations can flow round in a loop. That enables the network to do temporal processing and learn sequences, e.g., perform sequence recognition / reproduction or temporal association/prediction. Recurrent Neural Network architecture can have many different forms. One common type consists of a standard Multi-Layer Perceptron (MLP) plus added loops. These can exploit the powerful non-linear mapping capabilities of the MLP, and also have some form of memory. Others have more uniform structures, potentially with every neuron connected to all the others, and may also have stochastic activation functions. For simple architectures and deterministic activation functions, learning can be achieved using similar gradient descent procedures to those leading to the back-propagation algorithm for Feed-Forward Networks. When the activations are stochastic, simulated annealing approaches may be more appropriate. The general structure of RNN is given below Fig.7:

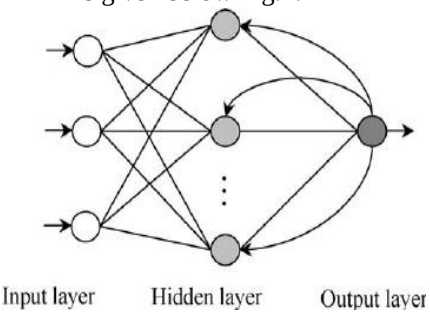


Fig.7. General structure of RNN.

The following are the most important features of recurrent networks.

- **Stability:** It concerns the boundedness over time of the network outputs, and the response of the network outputs to small changes (e.g., to the network inputs or weights).
- **Control:** It is concerns the possibility to control the dynamic behaviour. A Recurrent Neural Network is said to be controllable if an initial state is steerable to any desired state within a finite number of time steps.
- **Observation:** It is concerns the possibility to observe the results of the control applied. A recurrent network is said to be observable if the state of the network can be determined from a finite set of input/output measurements.

III. EXPERIMENTAL RESULTS

A. Dataset Collection

Table I. shows the voice samples used for training the neural network with each type of pathological voice. A total of 75 voice samples are used such as 15 voice samples from each type of voice pathologies.

TABLE I: Collection of data

SNO	AGE	Number of samples	Gender	Types of Voices
1	20-35	15	Female	Cyst (Pathology voice)
2	20-35	15	Male	Morbus Parkinson (Pathology voice)
3	20-35	15	Male, Female	Diplohonie (Pathology voice)
4	20-35	15	Male, Female	Fibrom (Pathology voice)
5	20-25	15	Male, Female	Vox-senillis (Pathology voice)

B. MFCC Feature Extraction

A total of 75 various pathological voice samples are used to the extract the MFCC features. The duration of the files is 0 to 2 secs each.

ip_mfcc ×

ip_mfcc <13x100 double>

	1	2	3	4	5	6	7	8	9	10
1	0.2337	0.8200	3.1623	2.7490	2.3849	1.1182	0.3160	0.4982	0.1826	0.47
2	0.6677	0.6081	0.9725	1.6056	0.9884	0.6097	0.2939	0.4211	0.2370	0.29
3	0.2240	0.5383	0.5846	1.3412	0.5840	0.6634	0.2023	0.4506	0.1523	0.30
4	0.1775	0.4408	0.6047	1.1978	0.4935	0.3159	0.2058	0.4931	0.1185	0.34
5	0.1946	0.4392	0.6338	1.1839	0.4397	0.4302	0.2694	0.3702	0.1105	0.44
6	0.2709	0.3118	0.5726	1.1271	0.3978	0.2904	0.2051	0.2664	0.1050	0.30
7	0.2076	0.2993	0.6254	0.9995	0.3866	0.2315	0.2251	0.3564	0.1128	0.37
8	0.2160	0.3362	0.5664	0.7097	0.4318	0.3042	0.2685	0.2491	0.1079	0.22
9	0.1814	0.2641	0.4984	0.6923	0.3890	0.3868	0.2186	0.3172	0.1029	0.23
10	0.1579	0.3030	0.5070	0.6772	0.3910	0.3882	0.2487	0.1671	0.0915	0.27
11	0.1414	0.2146	0.4386	0.6234	0.3671	0.2104	0.1859	0.3033	0.1194	0.20
12	0.1153	0.1959	0.4260	0.5636	0.3396	0.2456	0.1677	0.2090	0.1045	0.21
13	0.1252	0.1552	0.4336	0.4681	0.3666	0.3223	0.1449	0.1813	0.0973	0.25

Fig.8. MFCC features.

Figu.8. shows the MFCC features (13 dimensions) for each of the voice samples having 50 khz frequency.

C. Neural Network Classification

The MFCC features obtained is given to various types of neural networks like GRNN, MLPNN and RNN and the classification accuracy is given below,

TABLE II: Accuracy For GRNN With Different MFCC Coefficients

S. NO.	MFCC (No. of Coefficients)	Accuracy %
1	13	78.7
2	20	92
3	26	93.3
4	30	94.7

Table II. . shows the accuracy for GRNN with different MFCC coefficients. The table consists of 4 different trails with varying coefficients like 13, 20, 26 and 30. Their corresponding performance have been shown.

TABLE III: Accuracy for MLPNN Different MFCC Coefficients

S. NO.	MFCC (No. of Coefficients)	Accuracy %
1	13	58.7
2	20	78.1
3	26	81.4
4	30	84.0

Table III. shows the accuracy for MLPNN with different MFCC coefficients. The table consists of 4 different trails with varying coefficients like 13, 20, 26 and 30. Their corresponding performance have been shown.

TABLE IV: Accuracy for RNN Different MFCC Coefficients

S. NO.	MFCC (No. of Coefficients)	Accuracy %
1	13	62
2	20	66.7
3	26	57.0
4	30	62.6

Table IV. shows the accuracy for RNN with different MFCC coefficients. The table consists of 4 different trails with varying coefficients like 13, 20, 26 and 30. Their corresponding performance have been shown.

D. Performance Evaluation

The performance of the neural network is evaluated using the confusion matrix (error matrix). It shows the accuracy obtained for each output classes. We have five types of pathological classes. The confusion matrix with maximum accuracy using every neural network like GRNN, MLPNN and RNN is given below,

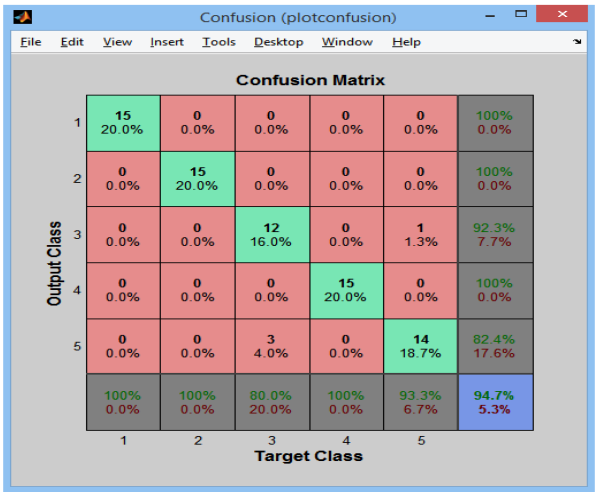


Fig.9. Confusion Matrix for GRNN.

Fig. 9. is the confusion matrix (error matrix) which is a plot for target class in the x-axis and output class in the y-axis. It can be used to assess the performance of a classifier. In this figure, the diagonal elements except the last one shows the percentage of correct classification by trained network for the five types of target classes. The overall accuracy for the classification of 5 classes of pathological voice for General Regression Neural Network is 94.7%. Fig.10. is the confusion matrix (error matrix) which is a plot for target class in the x-axis and output class in the y-axis. It can be used to assess the performance of a classifier. In this figure, the diagonal elements except the last one shows the percentage of correct classification by trained network for the five types of target classes. The overall accuracy for the classification of 5 classes of pathological voice for Recurrent Neural Network is 66.7%.

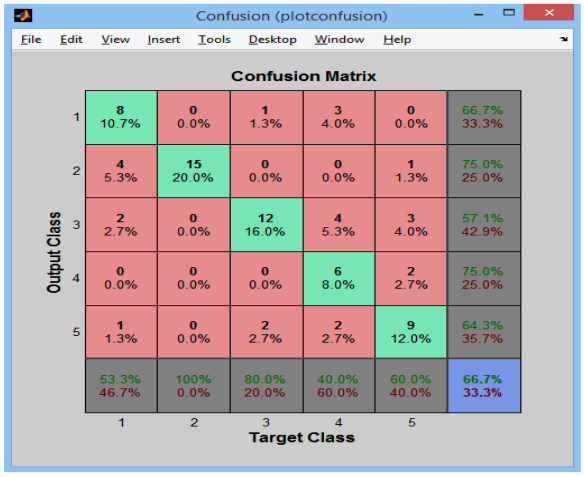


Fig.10. Confusion Matrix for RNN.

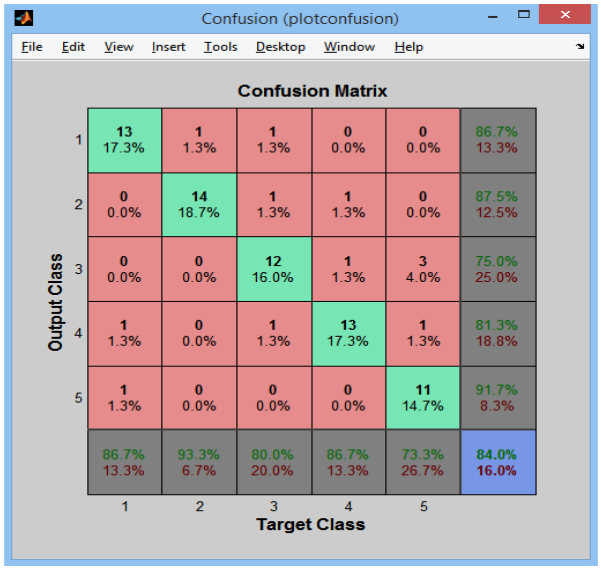


Fig.11. Confusion Matrix for MLPNN.

Fig.11. is the confusion matrix (error matrix) which is a plot for target class in the x-axis and output class in the y-axis. It can be used to assess the performance of a classifier. In this figure, the diagonal elements except the last one shows the percentage of correct classification by trained network for the five types of target classes. The overall accuracy for the classification of 5 classes of pathological voice for Multi Layer Perceptron Neural Network is 84.0%.

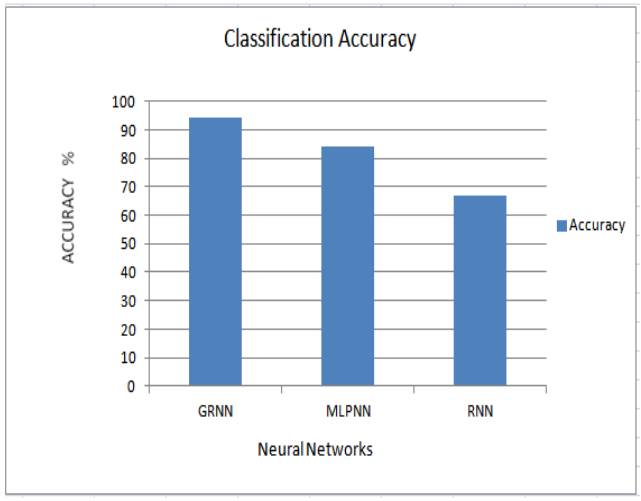
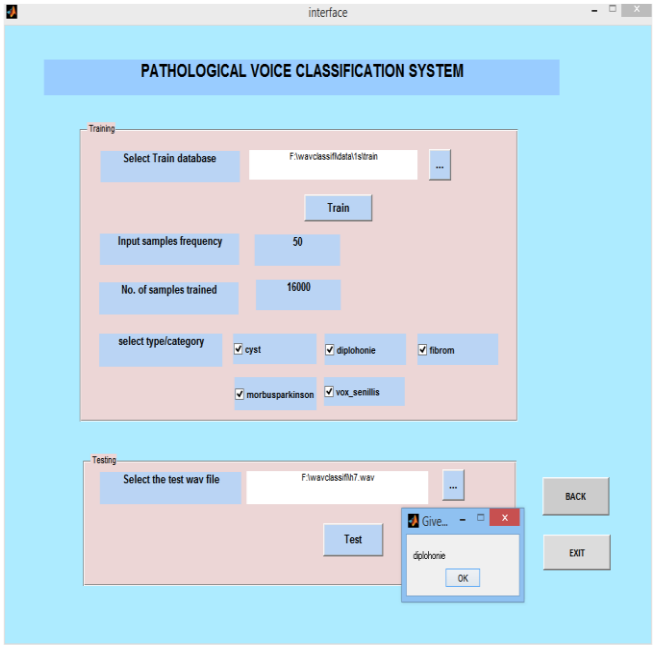


Fig. 12. Performance comparison for GRNN, MLPNN and RNN.

Fig.12. shows the General Regression Neural Network gives a maximum classification accuracy of 94.7%

E. Pathological Voice Classification System

The GUI window that has been created consists of two panels. First panel consists of the training part and the second panel consists of the testing part. The training part extracts the features from the given dataset and displays their frequency and the number of input samples used for training. The testing part detects and displays the type of pathological voice.



Figu13. Testing.

Fig.13. shows the testing results. The pathological voice is given as input in the form of a '.wav' file. The system detects it and displays the pathological voice type as "diplohonie". Similarly for the other types of pathological voices, the system detects and displays their corresponding pathological voice type.

IV. CONCLUSION

This work aims at developing an automated pathological voice classification system. This is an isolated speech classification system. Mel-Frequency Cepstral Coefficients are extracted from the audio recordings. These MFCC features are used as input for Artificial Neural Network model for classification. From the behaviour of the Artificial Neural Network (ANN), the classification of different types pathological voices has been analyzed. The various types of Artificial Neural Networks like General Regression Neural Network, Multi Layer Perceptron Neural Network, Recurrent Neural Network are used to classify the various types of pathological voices. The pathological voice classification system detects the different type of pathological voices such as cyst, diplohonie, morbus-parkinson, fibrom and vox-senillis. They can be used for pathology diagnosis in the speech production systems.

V. REFERENCES

[1]Abin Mathew George, Eva George, 'Detection of Voice Disguise by Various Disguising Factors', International Journal of Innovative Research in Computer and Communication Engineering, Vol 3, Issue 8, August 2015.

[2]Akansha Madan and Divya Gupta, 'Speech Feature Extraction and Classification: A Comparative Review', International Journal of Computer Applications, Vol. 90, No 9, March 2014.

[3]Arulmozhi P and Srinivasan V, 'Classification of Pathological Voice using Neural Network Model',

International Journal of Scientific Research and Technology, Vol 3, Issue 23, September 2014.

[4]Jose B. Trangol Curipe and Abel Herrera Camacho, 'Feature Extraction Using LPC-Residual and Mel Frequency Cepstral Coefficients in Forensic Speaker Recognition', International Journal of Computer and Electrical Engineering, Vol. 5, No. 1, February 2013.

[5]Lakshmi Kanaka Venkateswarlu Revada1, Vasantha Kumari Rambatla2 and Koti Verra Nagayya Ande, 'A Novel Approach to Speech Recognition by Using Generalized Regression Neural Networks', IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[6]Lotfi Salhi, Talbi Mourad and Adnene Cherif, 'Voice Disorders Identification Using Multilayer Neural Network', The International Arab Journal of Information Technology, Vol. 7, No. 2, April 2010.

[7]M. Hariharan, M. P. Paulraj, Sazali Yaacob, 'Time-Domain Features and Probabilistic Neural Network for the Detection of Vocal Fold Pathology', Malaysian Journal of Computer Science, Vol. 23(1), pp 60-67, December 2010.

[8]Namrata Dave, 'Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition', International Journal for Advance Research in Engineering and Technology, Vol. 1, Issue 6, July 2013 .

[9]Nawel SOUISSI, 'Artificial Neural Networks and Support Vector Machine for Voice Disorders Identification', (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, pp. 339-344, 2016.

[10]Pratik K. Kurzekar , Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, 'A Comparative Study of Feature Extraction Techniques for Speech Recognition System', International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, December 2014 .

[11]Saduf and Mohd Arif Wani, 'Comparative Study of Back Propagation Learning Algorithms for Neural Networks', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, December 2013.

[12]Salhi L, Talbi M, and Cherif A, 'Voice Disorders Identification Using Hybrid Approach: Wavelet Analysis and Multilayer Neural Networks', International Journal of Computer Applications, Vol. 47, No.13, June 2012.

[13]Shreya Narang1, Ms. Divya Gupta, 'Speech Feature Extraction Techniques: A Review', International Journal of Computer Science and Mobile Computing, Vol. 4, Issue 3, March 2015.

[14]Shruti and Bharti Chhabra, 'An Approach for Singer Identification Technique Using Artificial Neural Network', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 11, November 2015.

[15]Srinivasan V, Ramalingam V, and Arulmozhi P, 'Artificial Neural Network based pathological voice classification using MFCC features', International Journal of Science, Environment and Technology, Vol. 3, No 1, 2014.

[16]Srinivasan V, Ramalingam V and Arulmozhi P, 'Auto Associative Neural Network Based Classification of Pathological Voice using DWT and LPC Features', International Jounal of Scientific Engineering and Technology Research, Vol. 03, Issue. 03, March 2014.

[17]Taabish Gulzar, Anand Singh and Sandeep Sharma, 'Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks', International Journal of Computer Applications, Vol. 101, No.12, September 2014 .

[18]Vahid Majidnezhad, Igor Kheidorov, 'An ANN-based Method for Detecting Vocal Fold Pathology', International Journal of Computer Applications, Vol. 62, No. 7, January 2013.

[19] Vidushi Sharma, Sachin Rai and Anurag Dev, 'A Comprehensive Study of Artificial Neural Networks', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 10, October 2012.

[20]Vimala C, Radha V, 'Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words', (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , pp. 378-383, 2014.